Analysis of Large Genomic Data *in Silico*: The EPIC-Norfolk Study of Obesity

Jing Hua Zhao (赵京华)
MRC Epidemiology Unit, Cambridge, UK
jinghua.zhao@mrc-epid.cam.ac.uk
http://www.mrc-epid.cam.ac.uk/~jinghua.zhao

23/8/2007

MRC | Epidemiology Unit

Outline

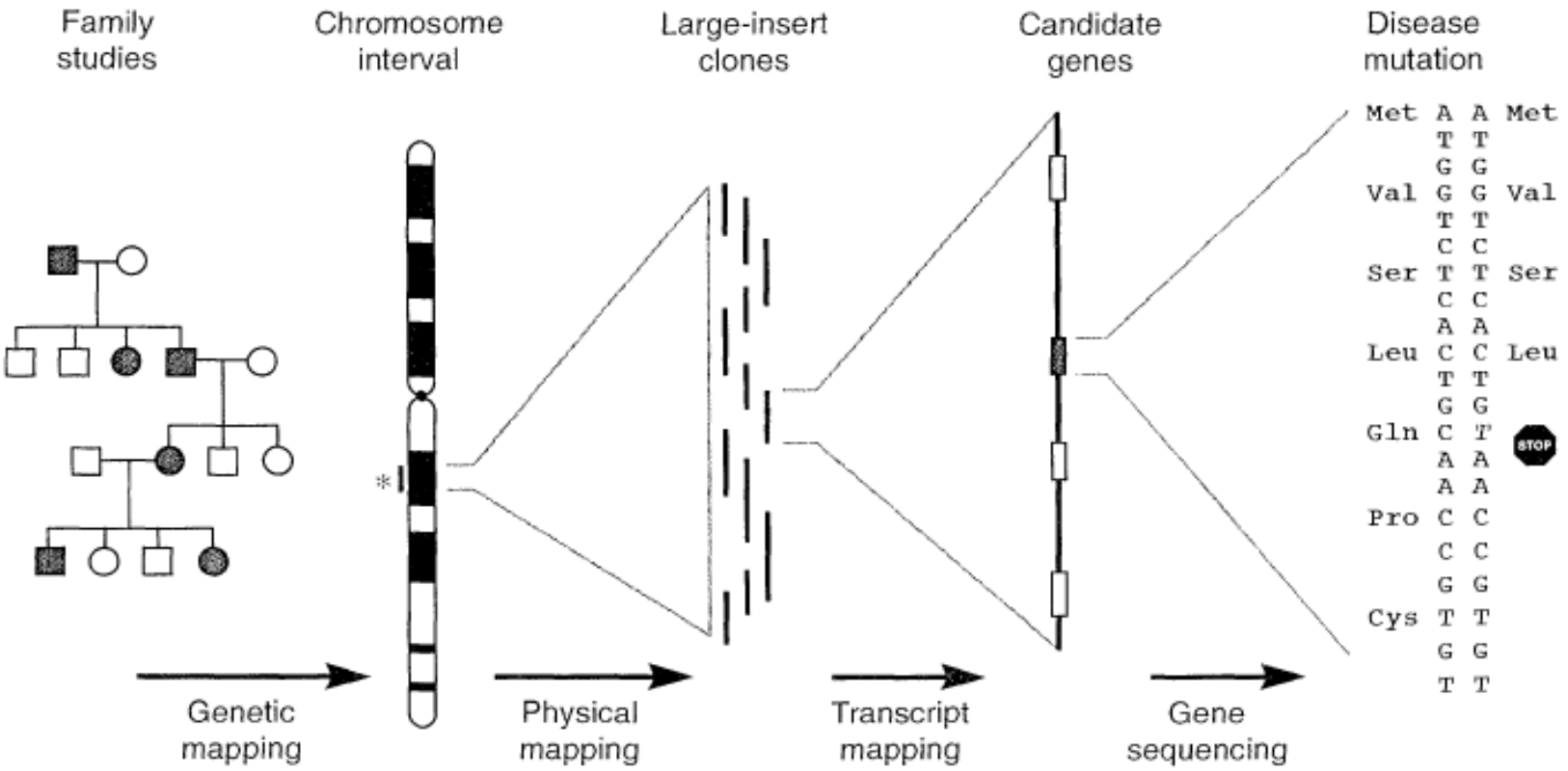
- Background
- Genomewide Association Studies (GWAS)
- Statistical/computational issues
- EPIC-Norfolk Study of Obesity
- SAS implementation
- Discussion
- Further information

Some Facts about Human Genome

- 23 (22 autosomal+1 sex) pairs of chromosomes
- $3x10^9$ DNA base pairs (bp), or 50,000~100,000 genes if 30,000bp per gene
- Human Genome Projects and HAPMAP projects (Couzin J. Science 296:1391-3, 2002; 310:601, 2005; Couzin J & Kaiser J. 316:822,2007)
- http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml and http://www.hapmap.org

Earlier Paradigm of Gene-Disease Association Study

- Lander ES & Schork NJ. Science 265:2037-2048,1994
- Schuler GD, et al. Science 274:540-546,1996



GWAS

- Owing to availability of large number of DNA variants called single-nucleotide polymorphisms (SNPs)
- Examples include Affymetrics 500k, Illumina 317k GeneChips
- Population-based or family-based samples
- A variety of problems associated with large data
- 10,000 individuals each with 1 million SNPs has a total of genotypes for 1GB storage if single byte variables are used. It is possible to pack SNP genotypes into bytes or bits. Moreover, the storage could be much reduced if data are organised by chromosomes.

Three Initiatives

- The hapmap project, a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals.
  - http://www.hapmap.org
- *The Wellcome Trust Case Control Consortium (WTCCC) was formed with a view to exploring the utility, design and analyses of GWA studies. It brought together over 50 research groups from the UK that are active in researching the genetics of common human diseases, with expertise ranging from clinical, through genotyping, to informatics and statistical analysis.*
  - http://www.wtccc.org
- The genetic association information network (GAIN, a public-private partnership of the Foundation for the National Institutes of Health, Inc., which include corporations, private foundations, advocacy groups, concerned individuals, and the National Institutes of Health. This initiative will take the next step in the search to understand the genetic factors influencing risk for complex human diseases.
  - http://www.fnih.org/GAIN2/home_new.shtml,

Example Studies

Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

A collaborative study of bipolar disorder, coronary artery disease, hypertension, type-1 diabetes, type-2 diabetes, crohn's disease, rheumatic arthritis

Statistical/Computational Perspectives

- Study designs
- Genotype calling, Hardy-Weinberg equilibrium tests, quality control, statistical modelling including data imputation
- Multiple-testing
- Meta-analysis
- Current data as snapshot of human evolutionary history which calls for such theory
- Data mining or machine learning

Solutions

- Algorithm evolution spans over 50 years
- Individually customised programs with limited capabilities (http://inkage.rockefeller.edu)
- Call for established systems (e.g. Reviews on Human Genomics) into play
  - SAS (http://www.sas.com, http://en.wikipedia.org/wiki/SAS)
  - Stata (http://www.stata.com)
  - S-PLUS (http://www.insightful.com)
  - R (http://www.r-project.org)

SAS

- An established, cross-platform, modular system (SAS/BASE, SAS/STAT, SAS/GRAPH, SAS/QC, SAS/ETS, SAS/OR, SAS/INSIGHT, SAS/IML, SAS/GENETICS, …) used in many government agencies, commercial companies and academic institutions
- Support for database such as Oracle and MySQL with ODBC facility
- Comprehensive modelling facilities in statistics, quality control and operations research
- Statistical graphics
- Powerful programming language and integrated development environments (IDEs) including grid computing, dedicated module and interface to other languages such as C
- Output to database and PDF/HTML/XML formats through the output delivery system (ODS)
- Excellent technical support (fully documented online) and a large user community

EPIC-Norfolk Study of Obesity

- Based on the European Prospective Study in Cancer of ~25,000 individuals at Norfolk (EPIC-Norfolk)
- A two-stage case-cohort design involving 2,500 subcohort and case samples at each stage, totalling ~7,000 individuals at both stages
- Study of obesity defined as body mass index (BMI=weight/height^2) over 30 which is associated with a range of health-related problems, e.g. cancer, diabetes and mental disorders
- Perlegen 250k + Affymetrics 500k + Illumina 317k GeneChips
- A number of collaborative cohorts

SAS Implementation for EPIC-Norfolk Study of Obesity

- EPIC-400 analysis (shown in the proceedings)
- Data sizes 400GB for the stage 1 Affy500k
- Data partition: Let *N*=# records, a group indicator is then created dynamically as

  *ceil(s/N*_n_)*

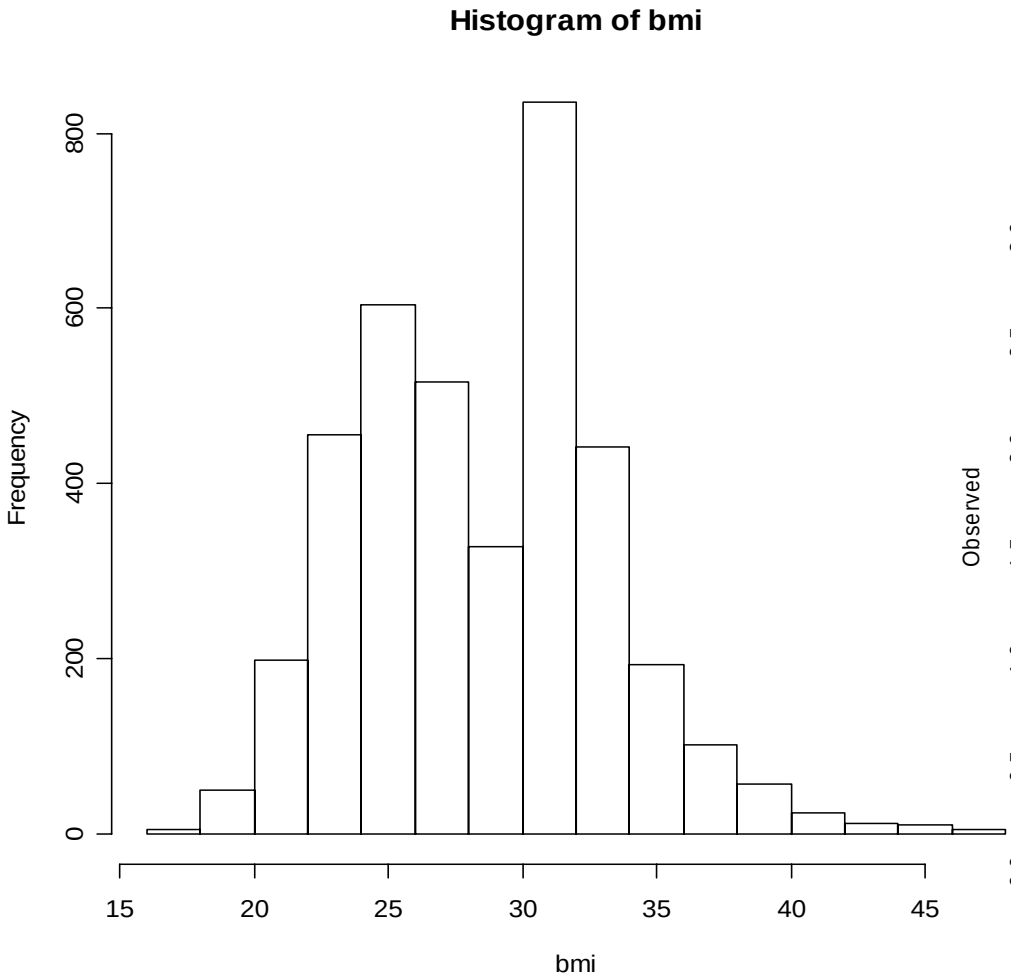
  where *s* is the number of partitions and _*n*_ is the running record number from SAS
- In future, grid computing will be used according to SAS/CONNECT
- Details from http://www.mrc-epid.cam.ac.uk/~jinghua.zhao

Findings

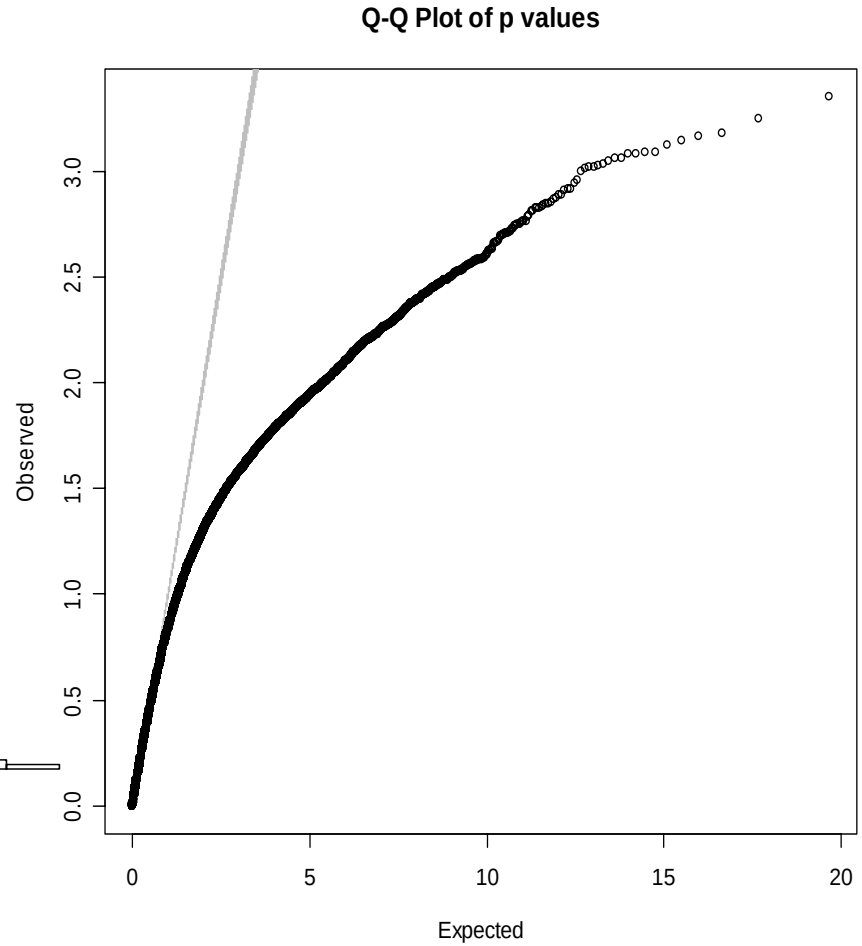
- SAS provides many features as an established traditional system appropriate for GWAS
- It is still preferable to incorporate faster algorithms
- At the moment it might be complementary to use both SAS and standalone programs to furnish specific aspects of the analysis
- In a long run standalone programs will be increasingly assimilated into systems such as SAS

A Common Variant in the *FTO* Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity

Timothy M. Frayling,[1,2]* Nicholas J. Timpson,[3,4]* Michael N. Weedon,[1,2]* Eleftheria Zeggini,[3,5]* Rachel M. Freathy,[1,2] Cecilia M. Lindgren,[3,5] John R. B. Perry,[1,2] Katherine S. Elliott,[3] Hana Lango,[1,2] Nigel W. Rayner,[3,5] Beverley Shields,[2] Lorna W. Harries,[2] Jeffrey C. Barrett,[3] Sian Ellard,[2,6] Christopher J. Groves,[5] Bridget Knight,[2] Ann-Marie Patch,[2,6] Andrew R. Ness,[7] Shah Ebrahim,[8] Debbie A. Lawlor,[9] Susan M. Ring,[9] Yoav Ben-Shlomo,[9] Marjo-Riitta Jarvelin,[10,11] Ulla Sovio,[10,11] Amanda J. Bennett,[5] David Melzer,[1,12] Luigi Ferrucci,[13] Ruth J. F. Loos,[14] Inês Barroso,[15] Nicholas J. Wareham,[14] Fredrik Karpe,[5] Katharine R. Owen,[5] Lon R. Cardon,[3] Mark Walker,[16] Graham A. Hitman,[17] Colin N. A. Palmer,[18] Alex S. F. Doney,[19] Andrew D. Morris,[19] George Davey Smith,[4] The Wellcome Trust Case Control Consortium,† Andrew T. Hattersley,[1,2]‡§ Mark I. McCarthy[3,5]‡



The BMI distribution in the EPIC-Norfolk study of obesity. The case-cohort sample is a combination of the sub-cohort sample and case sample which is truncated from the whole EPIC cohort at BMI=30.



Q-Q plot of p values from PROC QLIM

A Comparison with Standalone Programs

- Most implementations (e.g. PLINK: Purcell et al. *Am J Hum Genet,* 2007; EIGENSTRAT. Price et al. *Nat Genet.* 2006; SNPTEST: Marchini et al. *Nat Genet* 2007) have consolidated forms
  - Data management
  - Goodness of fit test (HWE) and association testing (Hotelling's T^2, etc)
  - Multivariate analysis (MVA, e.g., regression, cluster, correspondence, principal component, latent class, multidimensional scaling) and confirmatory analysis
  - Graphics
- However, it is preferable to have or tune in SAS (or indeed in other systems) independent compression/decompression and data-handling algorithms implemented
- The driving forces are that genetic data are both increasingly available and integrated, so that established systems are required and implementation involves a large number of collaborators

Challenges and Prospects

- This work as yet serves as an example
- In general, it is necessary to have the integration of three distinctive disciplines
  - Mathematical statistics, gene-gene interaction (GGI), gene-environment interaction (GEI), network analysis (Friedman N. Science 303:799-805,2004; Christakis et al. N Eng J Med 357:370-9, 2007)
  - Computer science
  - Biology, e.g. innovative elucidation of pathways
- The common issues in different disciplines possibly bear different names
- This collaborative endeavour will shed light on personalised medicine

Acknowledgements

- The ICIC2007 Programme Committee and Editors, anonymous reviewers for comments on the manuscript
- The MRC Epidemiology Unit
- The Wellcome Trust Sanger Institute
- The University of Cambridge

NHS

MRC | Epidemiology Unit

UNIVERSITY OF CAMBRIDGE

Institute of Metabolic Science

(http://www.ims.cam.ac.uk)